UNIVERSITY OF CAMBRIDGE

Miteyan Patel
Robinson College
mp781@cam.ac.uk

Computer Science Tripos - Part II Project Progress Report

Comparing Machine Learning Techniques for Mobility Graph Classification

February 1, 2018

supervised by
Dr. Sandra Servia

with Director of Studies
Prof. Alan Mycroft

and Overseers
Dr. Rafal Mantiuk & Dr. Ian Wassell

Project Schedule

My project involves comparing and evaluating supervised machine learning algorithms for classifying graphs. The graphs are composed of user's mobility patterns collected using their smartphone location data. The best of these will be then exported to an Android Application to make classifications locally on the device.

The principal components are the graph feature extractions, application of the machine learning algorithms to the graphs, evaluation of the machine learning algorithms, exporting to the Android Application.

My project is on schedule and one main feature needs to be implemented. The evaluation and selection of the best performing model to be imported onto the Android Application. This should not be too big of an issue since the rest of the project implementation has been completed and provides the framework for this completion. I am on track to finish this part of the project by the suggested date of February 23rd.

Work completed

- I have identified and extracted the relevant graph features from the graphs created and clustered from the Mobile Challenge Dataset[1] that are used for inputs into the supervised learning algorithms.
- Used machine learning algorithms from TensorFlow and Scikit-learn [2, 3]. These use, as input data, the relevant feature identified from the previous point or in the case of the CNN the graph edge lists. I have also implemented and tested my own logistic regression neural network in Java to compare performance.
- I have optimised the different learning algorithms for accuracy and other metrics.
- I have used the CNN algorithm proposed in [4] and modified the open source implementation[5] accordingly to classify the location graphs. These graphs have been previously created from the location dataset [1].
- I have implemented and tested a data clustering algorithm and an algorithm to construct a graph and its edge lists from the list of locations and times of an individual. Although this was initially indicated as an extension, it was necessary to increase the number of graphs and the quality of graph data and to try out different graph topologies.
- Adapted the Android Application provided that collected the location of the user safely to add clustering to group nearby locations to the same node in the graph. I also added graph construction to create graph edge lists from the clustered data to be used for the classification task.
- Tested major features and nearly written the first three chapters of the dissertation.

Work to complete

As mentioned earlier, I am yet to export a machine learning model to the Android application. I have partially implemented the Android application since it can cluster the close by locations recorded into the same cluster which represents a node in the mobility graph.

Apart from this, I have the extension work to complete apart from using and trying out different types of graphs, a part of the extension already completed.

Challenges faced

One of the earliest challenges I faced was improving and optimising the machine learning algorithms to best classify the graphs with the limited data available. Different data usage tricks and algorithms had to be used to maximise the effectiveness of the data and improve the accuracy in a domain not very documented or attempted before.

Another challenging aspect of my project has been working with the open source implementation of the graph CNN from the paper[4] as there was little documentation and I had not expected at the start of the project that it was not fully geared towards graph classification by itself.

References

[1] Kiukkonen, N., Blom, J., Dousse, O., Gatica-Perez, D., & Laurila, J. (2010). Towards rich mobile phone datasets: Lausanne data collection campaign. Proc. ICPS, Berlin.

[2] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., & Ghemawat, S. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.

[3] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake Vanderplas, Arnaud Joly, Brian Holt, Gal Varoquaux. API design for machine learning software: experiences from the scikit-learn project. arXiv:1309.0238.

[4] Defferrard, M., Bresson, X.,& Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. In Advances in Neural Information Processing Systems (pp. 3844-3852).

[5] Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering open source implementation: *github*